3η ΕΡΓΑΣΙΑ ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΠΛΑΤΩΝ ΚΑΡΑΓΕΩΡΓΗΣ 3180068
ΝΙΚΟΛΕΤΑ-ΚΛΕΙΩ ΠΑΤΑΤΣΗ 3180266

1ο Ερώτημα

Α) Με βάση τα φροντιστήρια και τις διαλέξεις του μαθήματος, αρχικά πρέπει να ελέγξουμε τις συνθήκες προκειμένου να συμπεράνουμε, εάν τα δεδομένα μας είναι κατάλληλα, για να εφαρμόσουμε τις μεθόδους της συμπερασματολογίας που γνωρίζουμε. Στην περίπτωση αυτή, είναι κατάλληλα γιατί έχουμε πάρει τα δεδομένα προφανώς από τυχαία δειγματοληψία, έχουμε πάνω απο 15 τιμές και δεν υπάρχουν σημαντικά outliers. Αρχικά, πρέπει να χρησιμοποιήσουμε τον τύπο «mean(x) + c(-1,1)*t*sd(x)/sqrt(n)». Από όλα τα στοιχεία του τύπου το μόνο που χρειάζεται περαιτέρω ανάλυση είναι το t, όπου είναι η τιμή της κατανομής και για να υπολογιστεί χρειάζεται ένα ποσοστημόριο και ο βαθμός ελευθερίας όπου είναι n-1, με n των αριθμό των δεδομένων. Δηλαδή, «t<- -qt(0.025, df=n-1)» στο t*. Το 0.025 το βρίσκουμε με την εξής λογική. Ψάχνουμε 95% πιθανότητα επομένως από 0 μέχρι 1 είναι 0.95 και άρα έξω από το ζητούμενο εμβαδόν έχουμε 0.05 πιθανότητα. Όμως η κανονική κατανομή είναι συμμετρική, επομένως το ποσοστημόριο είναι 0.05 διά 2 άρα 0.025. Στο mean(x) βάζουμε για x ένα σύνολο τιμών από άσσους και μηδενικά όπου οι άσσοι αντιπροσωπεύουν τις κορώνες και τα μηδενικά τα γράμματα. Το n προφανώς είναι 50. Το διάστημα εμπιστοσύνης που προκύπτει αν βάλουμε όλα τα παραπάνω στην R είναι 0.4383081 0.7216919.

Β) Το επίπεδο σημαντικότητας α, ορίζει την απόσταση που πρέπει να έχει ο δειγματικός μέσος από τη μηδενική υπόθεση, για να θεωρηθεί το αποτέλεσμα στατιστικά σημαντικό. Άρα, έστω H0: μ=μ0 το νόμισμα είναι δίκαιο και H1: μ!=μ0 το νόμισμα δεν είναι δίκαιο. Θα υπολογίσουμε το pvalue. Αρχικά πρέπει να βρούμε το $z = \hat{p} - p/\mathrm{sqrt}(p0(1\text{-}p0)/n)$. Το $\hat{p}$ είναι ίσο με 29/50, το p με 25/50, το p0 με 0.5 και το n με 50. Το αποτέλεσμα είναι z= 1.131371. Με αυτή την τιμή τρέχουμε στην R την εντολή 2* pnorm(-z) δηλαδή το 2-Φ(-Z) για να βρούμε το pvalue. Αυτό είναι ίσο με 0.257899 και άρα συμπεραίνουμε ότι είναι μεγαλύτερο από το 0.05 του επιπέδου σημαντικότητας και επομένως αποτυγχάνουμε να απορρίψουμε τη μηδενική υπόθεση. Αυτό είναι από άποψη λογικής σωστό καθώς 29 κορώνες σε 50 προσπάθειες είναι ένα λογικό νούμερο.

Γ) Τα δεδομένα προέρχονται από απλά τυχαία δείγματα που λήφθηκαν με ανεξάρτητο τρόπο από τους δυο πληθυσμούς. Θα εφαρμόσουμε τον τύπο n>= (z*^2) * p(1-p)/m^2. Έχουμε z* = 1.96, p=0.5 και m=0.01. Το αποτέλεσμα είναι 9604. Άρα με 9604 ρίψεις θα είχαμε τον ελάχιστο αριθμό ρίψεων που χρειάζονται για να μειώσουμε το περιθώριο λάθους σε μικρότερο του 1%.

2ο Ερώτημα

Πρώτα πρέπει να ελέγξουμε τις συνθήκες προκειμένου να συμπεράνουμε, εάν τα δεδομένα μας είναι κατάλληλα, για να εφαρμόσουμε τις μεθόδους της συμπερασματολογίας που γνωρίζουμε. Στην περίπτωση αυτή, είναι κατάλληλα γιατί τα δεδομένα προέρχονται από απλά τυχαία δείγματα που λήφθηκαν με ανεξάρτητο τρόπο από τους δυο πληθυσμούς και έχουμε ικανοποιητικό αριθμό επιτυχιών και αποτυχιών στα δείγματα. Το μέγεθος του πληθυσμού δεν περιλαμβάνεται στους παράγοντες του περιθωρίου σφάλματος. Αυτό σημαίνει, ότι ένα τυχαίο δείγμα μεγέθους 1000 ατόμων, ανεξάρτητα από το αν προέρχεται από έναν πληθυσμό μεγέθους 300 εκατομμυρίων ή 10 εκατομμυρίων, έχει το ίδιο περιθώριο σφάλματος το οποίο εξαρτάται μόνο από τα p̂, n, α. Μπορούμε να χρησιμοποιήσουμε το ίδιο δείγμα των 1100 ατόμων που χρησιμοποιήσαμε και στην Ελλάδα και να έχουμε αντίστοιχο περιθώριο σφάλματος 3%. Αυτό αποδεικνύεται αντίστοιχα, με τον τύπο που χρησιμοποιήσαμε στην προηγούμενη άσκηση n>= (z*^2) * p(1-p)/m^2 όπου αυτή τη φορά m=0.03. Το αποτέλεσμα είναι περίπου 1067. Άρα είτε με 1067 άτομα είτε με 1100 όπως στην περίπτωση της Ελλάδας, η δουλειά για τις δημοσκοπήσεις θα γίνει.

3ο Ερώτημα

Α) Ελέγχουμε τις συνθήκες προκειμένου να συμπεράνουμε, εάν τα δεδομένα μας είναι κατάλληλα για να εφαρμόσουμε τις μεθόδους της συμπερασματολογίας. Εδώ, είναι κατάλληλα γιατί τα δεδομένα προέρχονται από απλά τυχαία δείγματα που λήφθηκαν με ανεξάρτητο τρόπο από τους δυο πληθυσμούς και έχουμε ικανοποιητικό αριθμό επιτυχιών και αποτυχιών στα δείγματα. Ορίζουμε ως μηδενική υπόθεση H0: μ = μ0 την περίπτωση να υπάρχει συσχέτιση. Αντίστοιχα μ!=μ0 η περίπτωση να μην σχετίζονται. Θα εφαρμόσουμε το Welch Two Sample t-test. Παρακάτω βλέπουμε το αποτέλεσμα.

```
> t.test(real$FYLLO==0 & real$KAPNISTHS=="YES",real$FYLLO==1 & real$KAPNISTHS=="YES")

        Welch Two Sample t-test

data:  real$FYLLO == 0 & real$KAPNISTHS == "YES" and real$FYLLO == 1 & real$KAPNISTHS == "YES"
t = -0.43982, df = 117.63, p-value = 0.6609
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1834198  0.1167531
sample estimates:
mean of x mean of y
0.2000000 0.2333333

> |
```

Το pValue είναι 66% επομένως δεν μπορούμε σίγουρα να απορρίψουμε τη μηδενική υπόθεση.

Β) Υπολογίζουμε τα εξής. Πρώτα θα υπολογίσουμε το p1hat για τους άντρες, που είναι ίσο με 12/30, ακολούθως το p2hat για τις γυναίκες που είναι ίσο με 14/30. Το z* είναι ίσο με 1.96 αφού βρισκόμαστε στο 95% επίπεδο εμπιστοσύνης και έχω n1=n2=30. Άρα, τρέχω στην R: ci1<-p1hat-p2hat - 1.96*sqrt((p1hat*(1-p1hat))/n +(p2hat*(1-p2hat))/n) και ci2<-p1hat-p2hat + 1.96*sqrt((p1hat*(1-p1hat))/n +(p2hat*(1-p2hat))/n) με αποτέλεσμα το ακόλουθο διάστημα -0.3168743 0.183541.

Γ) Αφού περάσουμε τον πίνακα δεδομένων στην R, τρέχουμε την εντολή table(x$KAPNISTHS=="YES",x$FYLLO)->t και ακολούθως chisq.test(t,correct = FALSE). Αυτό μας επιστρέφει το ακόλουθο αποτέλεσμα.

```
> chisq.test(t,correct=FALSE)

        Pearson's Chi-squared test

data:  t
X-squared = 0.27149, df = 1, p-value = 0.6023

> |
```

Βλέπουμε ότι το pValue είναι 60% επομένως, δεν μπορούμε σε καμία περίπτωση να απορρίψουμε τη μηδενική υπόθεση αφού είναι πολύ μεγάλο. Από όσο γνωρίζουμε θα μπορούσαν και να σχετίζονται. Μετά φτιάχνω τον πίνακα συνάφειας όπως φαίνεται παρακάτω.

```
> prop.table(table(real$FYLLO,real$KAPNISTHS),margin=1)

           NO       YES
  0 0.6000000 0.4000000
  1 0.5333333 0.4666667
```

Δ) Το pValue από το Welch τεστ ήταν 0.6609 και του chi-square test 0.6023. Να σημειωθεί ότι χωρίς το correct=false το chi-square test ήταν 0.7945. Άρα, το Welch τεστ είναι πιο αυστηρό από το chi-square test εάν δεν χρησιμοποιήσουμε το Yates correction.

4ο Ερώτημα

Α) Πρώτα πρέπει να ελέγξουμε τις συνθήκες προκειμένου να συμπεράνουμε, εάν τα δεδομένα μας είναι κατάλληλα, για να εφαρμόσουμε τις μεθόδους της συμπερασματολογίας που γνωρίζουμε. Στην περίπτωση αυτή, είναι κατάλληλα γιατί τα δεδομένα προέρχονται από απλά τυχαία δείγματα που λήφθηκαν με ανεξάρτητο τρόπο από τους δυο πληθυσμούς και ο αριθμός του δείγματος των smarties είναι ικανοποιητικός. Έστω ότι έχουμε ένα σύνολο P με όλα τα smarties του κόσμου και παίρνουμε ένα δείγμα n=80. Σε αυτά τα 80 έχουμε 19 κόκκινα και 15 μπλε άρα έχουμε phat1 = 19/80 και phat2 = 15/80. Επίσης έχουμε phat = 34/160. Το 80 είναι και για τις δύο περιπτώσεις ίδιο γιατί ο πληθυσμός είναι κοινός. Θα υπολογίσουμε το z εφαρμόζοντας τον ανάλογο στατιστικό έλεγχο. Έστω μηδενική υπόθεση H0: μ=μ0 ότι τα κόκκινα και τα μπλε smarties είναι ίδια σε αριθμό. Αντίστοιχα, έστω Ηα: μ>μ0 όπου τα κόκκινα smarties είναι περισσότερα από τα μπλε. Έχουμε z= (phat1-phat2)/2*(phat(1-phat))/n. Το αποτέλεσμα είναι z= 0.773028. Έπειτα εφαρμόζουμε τον τύπο 1-pnorm(z) καθώς θέλουμε να δούμε αν είναι περισσότερα τα κόκκινα smarties. Το pvalue που προκύπτει είναι 0.2197529. Άρα δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση καθώς το pvalue είναι αρκετά μεγάλο.

Β) Το ποσοστό των κόκκινων smarties στο δείγμα μας είναι 19*100/80=23.75% και το 2009 ήταν 17.8%, άρα παρατηρούμε μία απόκλιση κοντά στο 6%. Στα μπλε smarties στο δείγμα μας

το ποσοστό είναι 15*100/80=18.75% και το 2009 ήταν 19.6%, άρα έχουμε μία μικρότερη απόκλιση γύρω στο 1%. Στα κίτρινα smarties έχουμε 16*100/80=20% και το 2009 ήταν 17.6% και επομένως παρατηρείται απόκλιση περίπου 2.5%. Επιπλέον, στα καφέ smarties έχουμε 22*100/80=27.5% και το ποσοστό του 2009 είναι 19.8% και η απόκλιση γύρω στο 7.5%. Τέλος, στα πράσινα smarties έχουμε ποσοστό ίσο με 8*100/80=10% και το 2009 είχαμε 25.2%, δίνοντας στην απόκλιση τη μεγαλύτερη της τιμή ως τώρα, καθώς αγγίζει το 15%. Καταλήγουμε, ότι ναι, έχει αλλάξει η κατανομή και το κάνει σαφές το παράδειγμα των πράσινων smarties, που το 2009 ήταν η πλειοψηφία, αλλά πλέον οι ενδείξεις μας από το δείγμα που έχουμε τώρα,μας οδηγούν στη σκέψη, ότι τα πράσινα smarties, είναι πλέον η απόλυτη μειοψηφία καθώς συγκεντρώνουν το μικρότερο ποσοστό.

Γ) Πρώτα πρέπει να ελέγξουμε τις συνθήκες προκειμένου να συμπεράνουμε, εάν τα δεδομένα μας είναι κατάλληλα, για να εφαρμόσουμε τις μεθόδους της συμπερασματολογίας που γνωρίζουμε. Στην περίπτωση αυτή, είναι κατάλληλα γιατί τα δεδομένα προέρχονται από απλά τυχαία δείγματα που λήφθηκαν με ανεξάρτητο τρόπο από τους δυο πληθυσμούς και ο αριθμός του δείγματος των smarties αλλά και των M&M's είναι ικανοποιητικός. Έστω μηδενική υπόθεση H0: μ=μ0 ότι έχουν ίδιες αναλογίες και αντίστοιχα έστω HA:μ!=μ0 δηλαδή ότι έχουν διαφορετικές αναλογίες. Θα κάνουμε έναν έλεγχο για κάθε χρώμα. Για το κόκκινο έχουμε phat1<- 19/80, phat2<-12/56 και phat<-31/136. Υπολογίζουμε το z και ακολούθως το pvalue που είναι 0.75. Με τον ίδιο τρόπο υπολογίζουμε και τα υπόλοιπα pvalue για τα άλλα χρώματα. Έχουμε:
Redpvalue = 0.75
Bluepvalue = 0.68
Greenpvalue = 0.83
Brownpvalue = 0.19
Yellowpvalue = 0.04

Από τα παραπάνω αποτελέσματα συμπεραίνουμε, ότι αν έχουμε ένα επίπεδο σημαντικότητας α=0.05, τότε η μόνη περίπτωση που απορρίπτουμε με ασφάλεια τη μηδενική υπόθεση, είναι για το κίτρινο χρώμα. Άρα γενικά με βάση τα δεδομένα που έχουμε η εκτίμηση που κάνουμε είναι ότι οι αναλογίες δεν διαφέρουν ιδιαίτερα.